Handling Structural Divergences and Recovering Dropped Arguments in a Korean/English Machine Translation System*

Chung-hye Han[1], Benoit Lavoie[2], Martha Palmer[1], Owen Rambow[3], Richard Kittredge[2], Tanya Korelsky[2], Nari Kim[4], and Myunghee Kim[2]

[1] Dept. of Computer and Information Sciences/IRCS
Univ. of Pennsylvania, Philadelphia, PA 19104, USA
{chunghye, mpalmer}@linc.cis.upenn.edu
[2] CoGenTex, Inc., Ithaca, NY 14850-1589, USA
{benoit, richard, tanya, myunghee}@cogentex.com
[3] ATT Labs-Research, B233, Florham Park, NJ 07932, USA
rambow@research.att.com
[4] Konan Technology, Inc., Seoul 135-090, Korea
nari@konantech.co.kr

Abstract. This paper describes an approach for handling structural divergences and recovering dropped arguments in an implemented Korean to English machine translation system. The approach relies on canonical predicate-argument structures (or dependency structures), which provide a suitable pivot representation for the handling of structural divergences and the recovery of dropped arguments. It can also be converted to and from the interface representations of many off-the-shelf parsers and generators.

1 Introduction

This paper describes an approach for handling structural divergences ([1, 3, 4, 8]) and recovering dropped arguments for Korean to English translation. Given that the two languages are very different from each other in structure, many challenging problems arise, demanding sophisticated linguistic modeling. The basic elements of our approach include:

- Transfer rules based on syntactic lexico-structural transfer ([8]);
- Conversion rules using a Korean predicate-argument lexicon for converting parsed syntactic structures produced by an off-the-shelf Korean parser ([12]) to the syntactic structures used for transfer;

* The work reported in this paper was supported by contract DAAD 17-99-C-0008 awarded by the Army Research Lab to CoGenTex, Inc., with the University of Pennsylvania as a subcontractor and NSF Grant - VerbNet, IIS 98-00658. Owen Rambow's contribution to this paper was made when he was with CoGenTex, Inc. and Nari Kim's contribution was made when she was a visiting researcher at IRCS, UPenn.

– Generation rules using an English realization lexicon for recovering dropped arguments after transfer.

The current implementation and processing of the transfer, conversion and generation rules is done uniformly, using a syntactic lexico-structural based framework ([5]). Declarative transformation specifications indicate how the lexemes and their relevant syntactic structures (essentially, their syntactic projection along with syntactic/semantic features) are mapped from one level to another. A similar approach was used in previous work for English to Arabic and English to French translations ([8, 9]).

The corpus for this project is a set of Korean/English parallel texts that consist of battle scenario message traffic and military language training manual. These contain information on typical military events such as troop movement, intelligence gathering, and equipment supplies, among others. Each half has roughly 50,000 word tokens, and 6000 sentences.

This paper is structured as follows. In section 2, we introduce some linguistic issues that pose problems for Korean/English MT. In section 3, we present a brief overview of the implemented system. Section 4 presents the linguistic knowledge bases used for conversion, transfer and argument recovery. We conclude with sections 5 and 6 with a brief comparison to different approaches in other MT systems (e.g., LCS and CCLINC) and a discussion of future work. Although our system handles transfer in both Korean-to-English and English-to-Korean directions, in this paper we mainly concentrate on the Korean-to-English direction for the sake of exposition.

2 Some Linguistic Issues in Korean/English Machine Translation

While English canonically has rigid subject-verb-object (SVO) order, Korean is a verb-final language with free word order. For instance, ditransitive sentences in English have 'subject-verb-indirect object-direct object' order, as shown in the target sentence in Table 1. The corresponding Korean sentence can have 'direct object-indirect object-subject-verb' order, as shown in the source sentence in Table 1.[1] In our system, the grammatical functions of argument NPs are identified by the use of Yoon's Korean parser and conversion rules using the predicate-argument lexicon.

Unlike English, argument NPs can be deleted in Korean. For instance, in the source sentence in Table 2, which is a conditional sentence, the subject NP in the *if*-clause has been deleted and the subject NP and the object NP in the main clause have been deleted. Ideally, all the missing arguments should be identified in the output as in the target sentence in Table 2. With the addition of a discourse component, the references of the missing arguments can be restored.

[1] Korean examples in this paper are romanized for convenience.

SOURCE:	chuka	kongkwupmul-eul	103	ceonwiciweontaetae-eke	saryeongpu-ka	cueossta.
GLOSS:	additional	supply-Acc	103	FSB-Dat	headquarter-Nom	gave
TARGET:	Headquarters gave 103rd FSB additional supplies.					
OUTPUT:	Headquarters gave an additional supply to a 103 forward support battalion.					

Table 1. Word Order

In our system, the dropped arguments are recovered for the English translation output using English generation rules.[2]

SOURCE:	IBP	hwail-eul	keomsaekhaci	moshaess-tamyeon	cikeum	tasi	ponaekessta.
GLOSS:	IBP	file-Acc	retrieve	could_not-if	now	again	will_send
TARGET:	if (NP1) could not retrieve IBP file, (NP2) will send (NP3) again now.						
OUTPUT:	If one can not retrieve an IBP file, one will send it again now.						

Table 2. Dropped Arguments and Morphology

In addition to word order difference and dropped argument recovery, there are many transfer issues that arise from structural divergences, some of which will be presented in section 4.3.

3 Overview of the System

Figure 1 illustrates the major transformation steps in our system. Korean or English sentences (depending on what the source language is) are first parsed. The parser output is then reformatted and converted into *Deep Syntactic Structure* (DSyntS) based on Meaning Text Theory (MTT) ([7]) (See below for more details). These Korean or English DSyntSs are then transferred respectively into English or Korean DSyntS (depending on what the target language is) that are finally realized as English or Korean sentences.

The DSyntS representations are composed of nodes labeled by lexemes which correspond to meaning-bearing words (nouns, verbs, adjectives, adverbs) and directed arcs with dependency relation labels. The subject is labeled as 'I', the direct object as 'II' and the indirect object as 'III'; label 'ATTR' covers all adjuncts. Function words such as determiners, semantically empty auxiliary verbs,

[2] As the anonymous reviewer correctly pointed out, the reference of *one* in the antecedent clause and the reference of *one* in the consequent clause must be different.

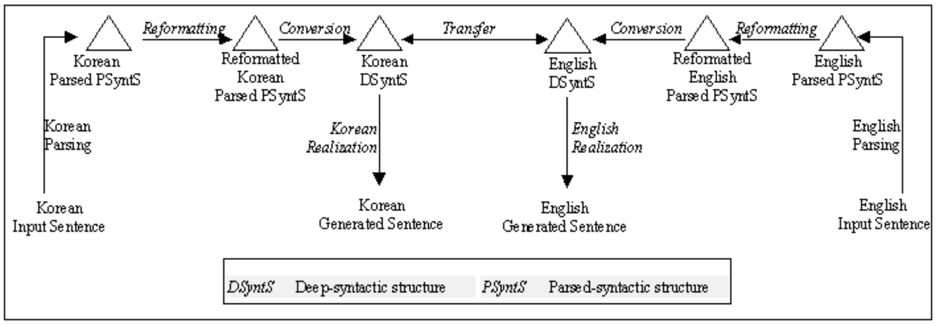

Fig. 1. Main Translation Steps

and grammatical morphology are represented through features on the node labels. This level of representation is well suited to MT since it abstracts away from superficial grammatical differences between languages, such as linear order and the usage of function words. For the sake of illustration, a DSyntS representation corresponding to the sentence *John often eats beans* is given in tree format below.

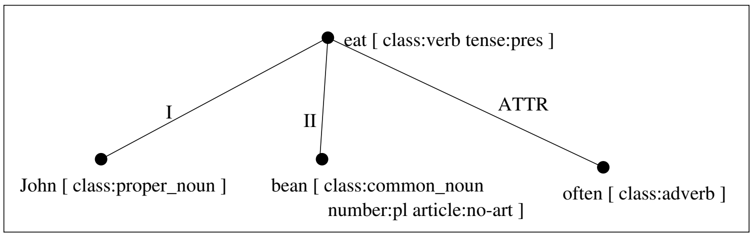

Fig. 2. DSyntS for *John often eats beans*

In our system, the conversion, transfer and realization is done uniformly via lexico-structural processing ([5]). The Korean parsing is done using Yoon's statistical Korean dependency parser ([12]), the English parsing is done using the Collins parser ([2]) and the Korean and English realization is done using RealPro ([6]).

4 Linguistic Knowledge Base

4.1 Predicate-Argument Lexicon

The predicate-argument lexicon contains subcategorization information for verbs and adjectives. The types of arguments (actants) include subjects (NP_0), direct objects (NP_1), indirect objects (NP_2), sentential complements (S_1) and optional arguments.

Example entries in the Korean predicate-argument lexicon are illustrated graphically in Figure 3. NP arguments are listed with case or adverbial postpositions as features: e.g., [case:] or [adv-case:]. Case postpositions include nominative and accusative case inflections and adverbial postpositions include those inflections that roughly correspond to English prepositions: e.g. {e-Ke} ('to'), {Ro} ('to'), {Kwa} ('with'), {e-Seo} ('from'). Sentential complements are listed with the relevant verbal inflectional morphology as a feature [mode-string:]. Actants with an asterisk (e.g., NP_1*) are optional arguments, and they count as arguments only when they are present in the sentence. In contrast, actants without an asterisk are obligatory arguments, and when they are missing from sentences, they are counted as dropped arguments. Example entries in the Korean predicate-argument lexicon are illustrated graphically in Figure 3.

In principle, all arguments are syntactically optional in Korean given that they can be dropped in the appropriate discourse context. Having said this, what we mean by optional/obligatory arguments are those that are optional/obligatory in the predicate-argument structure. Under this definition, for instance, in *John left for Paris*, *John* is an obligatory argument, but *for Paris* is an optional argument.

As will be described in Section 4.2, the Korean predicate-argument lexicon is used as a guide for making argument/adjunct distinctions in the DSyntS representations which are the input to the transfer component. The English predicate argument lexicon plays an important role in recovering arguments in English translation output when the corresponding input Korean sentence has dropped arguments, as will be discussed in section 4.4.

4.2 Conversion Rules

As mentioned in section 3, the source structures used for the transfer consist of MTT-based DSyntS. In our previous system we were able to utilize off-the-shelf English parsers and convert their output to our transfer lexicon requirements ([9]). The same approach has worked here, this time with a pre-existing Korean parser ([12]). This parser assigns dependencies between two words using lexical association values estimated on the basis of co-occurrence data extracted from a 30 million word corpus. The co-occurrence data consist of pairs of nouns for compound noun analysis, and triplets of a verb, an associated noun and the postposition on the noun for dependency analysis of verbs and nouns.

Although Yoon's Korean parser was not designed with DSyntS in mind, the generic dependency structure it produces is often isomorphic to the corresponding DSyntS. When it is not isomorphic, lexico-structural transformations can be

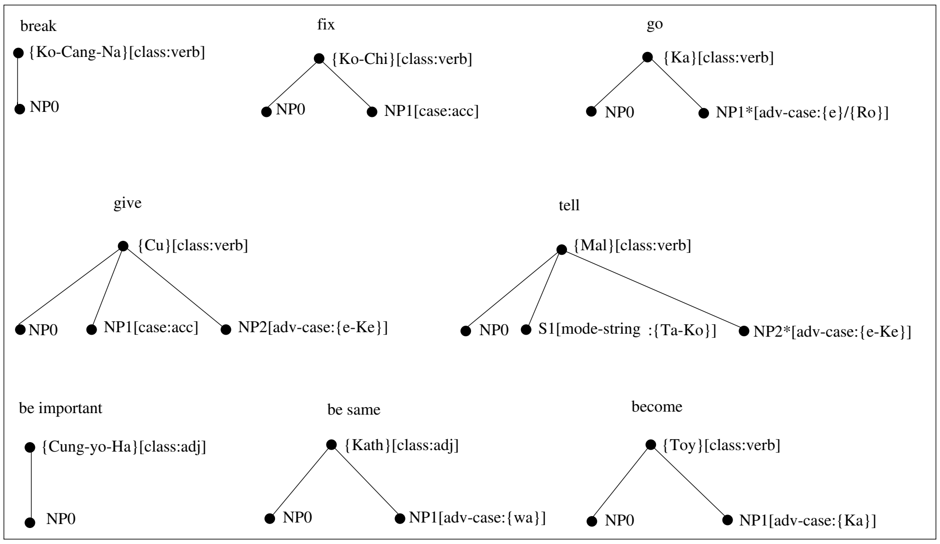

Fig. 3. Predicate-argument Lexicon for Korean

used to fix the discrepancies. The tree on the left in Figure 4 illustrates Yoon's parser output for the source sentence in Table 2 formatted using a tree notation. The tree on the right in Figure 4 illustrates the corresponding DSyntS used for transfer.

The transformation, or conversion, necessary to produce the DSyntS from the Korean parser output illustrated in Figure 4 takes place in three separate stages:

– *Rewriting feature labels*: Where our system requires different feature labels, preprocessing rules such as those in Figure 5 simply replace the feature 'ppca:{Reul}' with 'case:acc'.
– *Making dependency relationships more explicit*: In Figure 4, Yoon's parser only specifies one relation (the relation 'OBJ' between '{Keom-Saek-Ha}' and '{Hwa-il}'). The Korean predicate-argument lexicon (defined in section 4.1) is used as a guide for more explicit dependency relationships. The rule in Figure 6 sets the dependency relation between 'Keom-Saek-Ha' and 'X' to 'II' if 'X' has accusative case.
– *Promoting features to lexemes and vice versa*: Some of the features found in Yoon's Korean parser are represented as lexemes in the corresponding DSyntS. In Figure 7, the features 'enco2:{Ta-Myeon}' and 'ax:{Mos-Ha}' in Yoon's parser output are transformed to lexemes '{Ta-Myeon}' and '{Mos-Ha}' in the corresponding DSyntS. The rule in Figure 7 promotes the feature 'enco2:{Ta-Myeon}' to a lexeme. Predicate-specific lexico-structural grammar rules are used to map Yoon's argument structure onto ours.

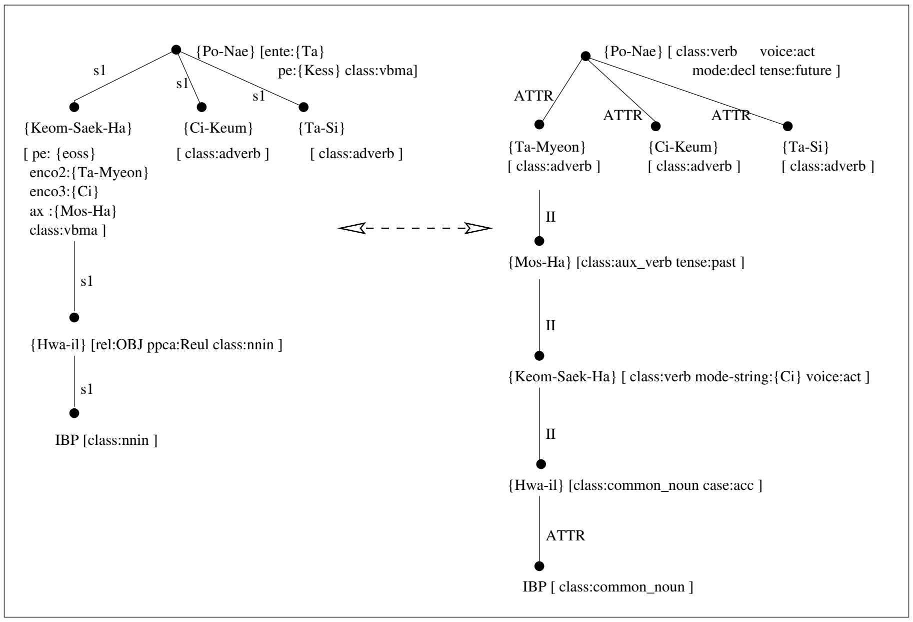

Fig. 4. Conversion from Reformatted Korean Parser Output to DSyntS

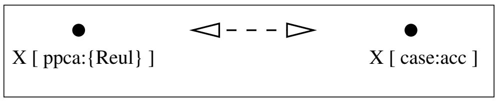

Fig. 5. Rewriting Feature Labels

This conversion process has thus provided the additional argument/adjunct distinctions that allow the parser output to be matched against our transfer lexicon, by referencing the Korean predicate-argument lexicon.

4.3 Transfer Lexicon

The transfer formalism is based on DSyntS grammars that are independently motivated by source and target languages, and was previously used for transfer from English to French and English to Arabic ([8]). It relates DSyntS subtrees, anchored by lexemes of different languages, with projections that represent a context in which the source language lexeme is translated into the target language lexeme. Transfer is carried out by replacing a subtree in the source language DSyntS with another to which it is linked in the target language DSyntS. In the simplest case, the related subtrees are reduced to a single node: the root of the

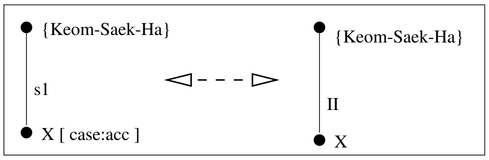

Fig. 6. Identifying the Dependency Relation

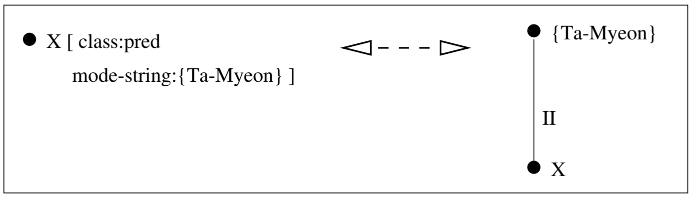

Fig. 7. Promoting Features to Lexemes and Vice Versa

tree. The following example shows a relation between the Korean verb $\{Po\text{-}Nae\}$ and the corresponding English verb *send*.

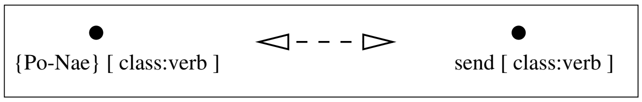

Fig. 8. Transferring Lexemes Directly

Additional contextual information is not required in this case since the two verbs share a common subcategorization frame. When applying such a rule for transfer, the nodes that are not present in the rule will remain unchanged after application of the rule. The target language realization grammars ensure that the proper word order for the target language is followed.

Multi-Word Transfer When the translation of a lexeme (or a group of lexemes) results in a syntactically divergent structure in the target language, this divergence is represented in the transfer lexicon by including contextual information in the related subtrees.

For instance, a predicative adjective in Korean translates to copular *be* and the corresponding adjective in English: e.g., $\{Cak\text{-}Ta\} \leftrightarrow$ *be small*. Such divergence necessitates a transfer rule that relates a single node anchored to one lexeme and a subtree with more than one node. The rule given in Figure 9 handles the transfer between predicative adjectives in Korean and English.

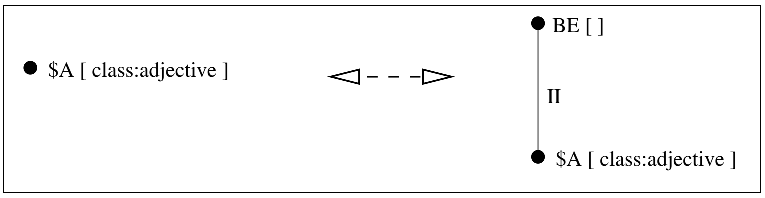

Fig. 9. Transferring Predicative Adjectives

Another example of syntactic divergence has to do with cases in which an inflection in one language translates to a lexical item in another language. For example, in Korean, the main verb in complement clauses has a verbal inflection which corresponds to the complementizer *that* in English. Our transfer rule that handles this divergence along with an example translation that uses this rule are given in Figure 10. We represent the verbal inflection as a feature [mode-string:Ta-Ko] on the verb node in the subordinate clause ($V2) and this node maps onto the corresponding verb and complementizer *that* in English.

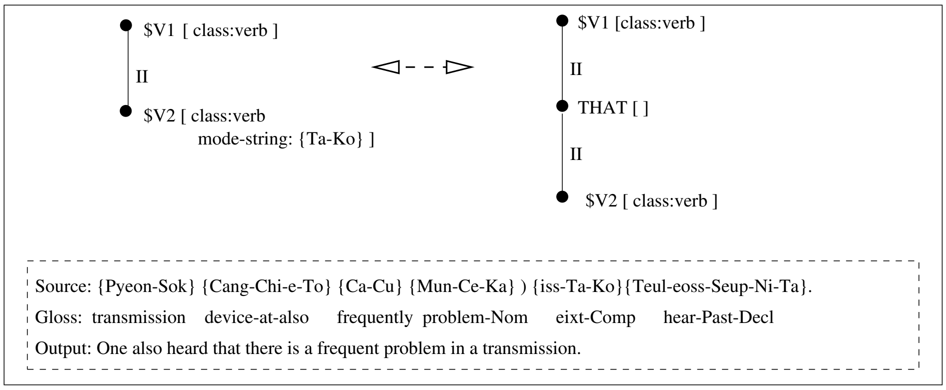

Fig. 10. Transferring Complementizer Inflection

An example of a more complicated structural divergence involves transferring a Korean complex NP whose head noun is lexicalized as an auxiliary noun {*Keos*} in the context of a copular to an English *to*-infinitive.[3] Our transfer rule that handles this divergence along with an example translation using this rule are given in Figure 11.

[3] Nouns that cannot stand alone are called *auxiliary nouns*. They must be modified by a clause or other nouns.

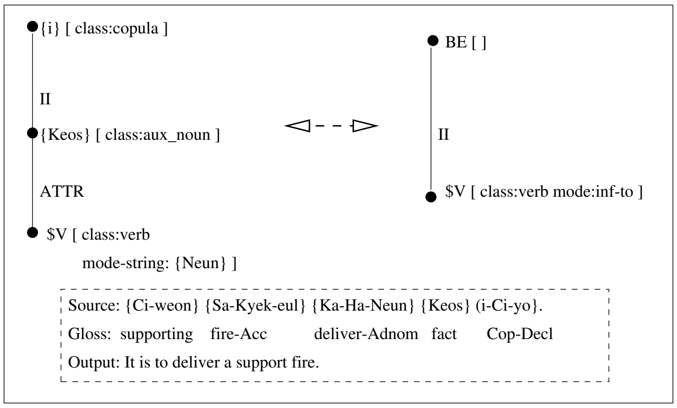

Fig. 11. Transferring Complex NP with Auxiliary Noun

Transfer rules can have variables (e.g., \$V1, \$V2 in Figure 10) instead of lexemes as nodes, allowing generalization of the rule application. Moreover, constraints on rule application can be introduced in the subtrees by means of features on the nodes. For instance, the feature [mode-string:{Neun}] in the transfer rule represented in Figure 11 restricts the rule application to source sentences containing a verb with inflectional morphology {-*Neun*}, which is an adnominal morpheme.

4.4 Argument Recovery Rules

In Korean, argument NPs can be dropped. When translating from Korean to English, the dropped arguments must be recovered in order to obtain grammatical English sentences. It is generally assumed that accurate translation of dropped arguments requires a discourse model. However, what type of discourse model is needed is not yet well understood. Our current translation model for Korean to English is based only on individual sentences and does not use a discourse model. Instead, we recover generic cases of dropped arguments by adding default pronouns for missing arguments using grammatical and lexical knowledge. For example, in Table 2, the three pronouns in the English translation, missing in the Korean sentence, have been recovered using only English grammatical and lexical knowledge: *If* **one** *can not retrieve an IBP file,* **one** *will send* **it** *again now.*

The recovery of dropped generic arguments is performed just before English realization, by preprocessing the English DSyntS obtained from the transfer of the corresponding Korean DSyntS. Figure 12 illustrates the result of the argument recovery processing applied to the English DSyntS generated after the transfer rules have applied to the example sentence in Table 2. Once again, the predicate-argument lexicon can be used as a reference to indicate the type of missing argument.

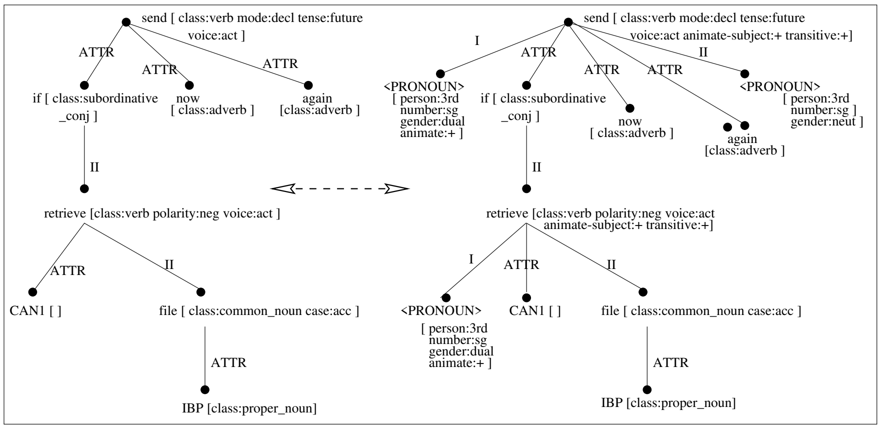

Fig. 12. Argument Recovery Transformation

Providing a filter for the argument is performed by using the following types of transformations, which currently do not take into account anaphoric dependencies:

– *Insertion of Missing Actant I*: This transformation involves adding, if missing, a 3rd person singular pronoun as actant I to a verb with mood indicative. Figure 13 illustrates a general case where a missing actant I can be recovered. Such a transformation is used in Figure 12 to add the actant I of ‘retrieve’ and ’send’ (*If* **one/it** *can not retrieve ...* **one/it** *will send ...*). (Another rule determines whether or not the pronouns must be animate.)

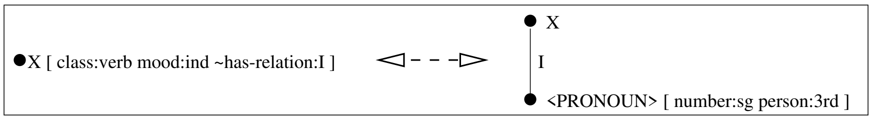

Fig. 13. Recovering Missing Actant I

– *Insertion of Missing Actant II*: This transformation involves adding, if missing, a 3rd person singular pronoun as actant II to a transitive verb of indicative mood and active voice. Figure 14 illustrates how a missing actant II can be recovered. Such a transformation is used in Figure 12 in order to add the actant II of ’send’ (*... one will send* **it** *...*).
– *Determining Whether Pronouns are Animate or Not*: This transformation involves setting the animate feature of a pronoun in the subject position. Figure 15 illustrates how the animate feature associated with a recovered

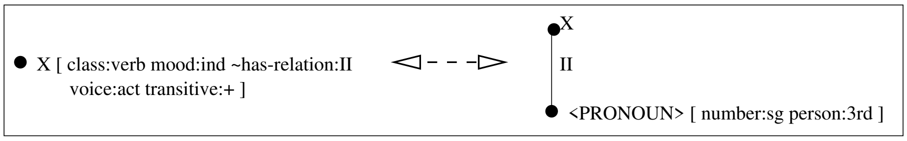

Fig. 14. Recovering Missing Actant II

pronoun can be determined in general. The transformation rule indicates how by unification, the value '?A' of the feature 'animate-subject:?A' assigned to a verb is passed to the pronoun. Such a transformation is used in Figure 12 to assign the feature 'animate:+' to the actant I of 'retrieve' and 'send' (*If* **one** *can not retrieve ...* **one** *will send ...*).

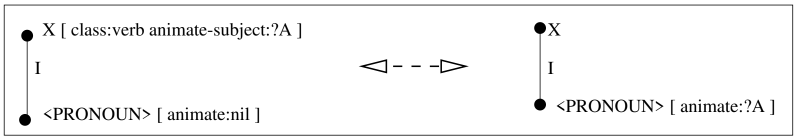

Fig. 15. Recovering of Animacy

4.5 Coverage

The current coverage of our linguistic knowledge base can be summarized as follows. Both the Korean and English predicate-argument lexicons cover 100% of the verbs and adjectives from our corpus. The transfer lexicon covers 100% of the word-to-word mappings, and 20% of these have currently been enriched for structural mapping and evaluated for the quality of the translations they produce. We should have 80% coverage by the end of August. In addition, we have a 20,000+ word bilingual lexicon for Korean/English provided by Systran, which contains subcategorization frame information for Korean, and word-to-word mappings to English. We will be enriching the Systran verb and adjective entries with the type of argument links that have been presented here.

The coverage for the Conversion Rules and Argument Recovery Rules is 100% of the small base set we tested for transfer (representing 20% of the corpus), although extension of the rules will be needed to cover the rest of the corpus. Since we have a parallel corpus, we have a Gold Standard for our target translations, and consider the translation is of suitable quality if it has the correct predicate argument structure, the expected lexemes and is grammatical.

5 Comparison to Other Approaches

A crucial part of an interlingua-based approach to MT, such as the LCS approach at the University of Maryland ([3]), and CCLINC at MIT's Lincoln Labs ([10]), is the mapping of a source language sentence to a language-independent intermediate representation, which serves as the basis for generating an output in the target language. The abstract interlingua has been argued to facilitate the development of a multilingual system, and to facilitate the handling of structural divergences. However, a disadvantage of the interlingua approach is the difficulty of reaching a consensus on criteria for truly language-independent representations. It is also necessary to develop special purpose language-specific parsers that can map the source language sentence onto the appropriate interlingua representation.

With our approach, there is no extra level of representation that mediates between source and target sentences. Instead, the source dependency parse tree is directly mapped to the target dependency tree. However, by basing our transfer rules on canonical predicate-argument structures with feature specifications, we provide a level of representation that can still capture the same structural divergences and many of the semantic generalizations traditionally associated with interlinguas ([8]). Our predicate-argument structures are in fact quite similar to the CCLINC "semantic frames," although somewhat flatter than LCS representations ([8]). In addition, we can much more readily utilize off-the-shelf parsers, as described above, and can also exploit statistical techniques for analyzing corpora and finding automatic alignments between parallel words and phrases ([9]). This provides the basis for our current experiments with the extraction of transfer lexicons from annotated bilingual corpora by decomposing tree structures into lexicalized subtrees ([11]).

6 Conclusion

We have described an approach for handling structural divergences and recovering dropped arguments for Korean to English machine translation. The common reliance on canonical predicate argument structure representations of lexical items provides the basis for our treatment of structural divergences and recovery of dropped arguments.

Our future plans include the development of a Korean TreeBank for our corpus which will be based on the hand corrected output of our Korean parser. We will also use the Korean and English parsers to help us construct a parallel, syntactically annotated corpus for automatic extraction of our lexico-structure transfer rules. The explicit annotation of empty arguments as well as the incorporation of a discourse model will allow a more principled recovery of implicit arguments.

Finally, we will extend the conversion rules and argument recovery rules to better cover the corpus, and enhance the machinery to improve the processing with larger knowledge bases.

References

1. Abeillé, A., Y. Schabes, and A. K. Joshi (1990) "Using Lexicalized Tags for Machine Translation," in *Proceedings of the International Conference on Computational Linguistics (COLING '90)*, Helsinki, Finland.
2. Collins, M. (1997) "Three Generative, Lexicalized Models for Statistical Parsing," in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, Madrid, Spain.
3. Dorr, B. J. (1993) *Machine Translation: a View from the Lexicon*, MIT Press, Boston, Mass.
4. Dorr, B. J. (1994) "Machine translation divergences: a formal description and proposed solution," *Computational Linguistics* 20:4, 597–635.
5. Lavoie, B., R. Kittredge, T. Korelsky, and O. Rambow (2000) "A Framework for MT and Multilingual NLG Systems Based on Uniform Lexico-Structural Processing," in *Proceedings of ANLP/NAACL 2000*, Seattle, Washington.
6. Lavoie, B., and O. Rambow (1997) "A Fast and Portable Realizer for Text Generation Systems," in *Proceedings of the Conference on Applied Natural Language Processing (ANLP'97)*, Washington, DC.
7. Mel'čuk, I. A. (1988) *Dependency Syntax: Theory and Practice*, State University of New York Press, New York.
8. Nasr, A., O. Rambow, M. Palmer, and J. Rosenzweig (1997) "Enriching Lexical Transfer with Cross-Linguistic Semantic Features," in *Proceedings of the Interlingua Workshop at the MT Summit*, San Diego, California.
9. Palmer, M., O. Rambow, and A. Nasr (1998) "Rapid prototyping of domain-specific machine translation system," In *Proceedings of AMTA-98*. Langhorne, PA, October.
10. Weinstein, C. J., Y. S. Lee, , S. Seneff, D. R. Carlson, B. Carlson, J. T. Lynch, J. T. Hwang, and L. C. Kukolich (1997) "Automated English/Korean Translation for Enhanced Coalition Communications," *Lincoln Laboratory Journal* 10:1, 35–60.
11. Xia, F. (1999) "Extracting Tree Adjoining Grammars from Bracketed Corpora," in *Proceedings of 5th Natural Language Processing Pacific Rim Symposium (NLPRS-99)*, Beijing, China.
12. Yoon, J., S. Kim, and M. Song (1997) "New Parsing Method Using Global Association Table," In *Proceedings of International Workshop on Parsing Technology.*